Abstract

We give a review on the rigorous results concerning the storage capacity of the Hopfield model. We distinguish between two different concepts of storage both of them guided by the idea that the retrieval dynamics is a Monte–Carlo dynamics (possibly at zero temperature). We recall the results of McEliece et al. [MPRV87] as well as those by Newman [N88] for the storage capacity of the Hopfield model with unbiased i.i.d. patterns and comprehend some recent development concerning the Hopfield model with semantically correlated or biased patterns.

1 Introduction and Two Concepts of Storage Capacity

Let us recall that one of the most important motivations to study the Hopfield model has always been that it can be regarded as one of the central and easiest models of a neural network and that it exhibits certain phenomena considered as the most important advantages of neural networks over ordinary computers. Especially, when considering the memory aspects of the Hopfield model the memory is diffused (in contrast to the localized computer memory) and content-addressable such that even strongly noised data can be successfully retrieved. Hence we may regard the Hopfield model as a toy model for modelling brain functions.

In this context the most natural question to ask is how many patterns the Hopfield model can store and how the maximum number of stored patterns scales with the number of neurons N. Already numerical investigations by Hopfield [Ho82] suggest that there is a critical value $\alpha_c \sim 0.14$ such that the Hopfield model can store less than $\alpha_c N$ patterns, if small errors are tolerated. This finding has been supported (with a similar value for α_c) by the non-rigorous analysis in [AGS87].

Before we give a mathematical analysis of the storage capacity of the Hopfield model we first have to briefly explain the two different concepts of storage we are dealing with on a technical level.
To this end let us first recall the definition of the Hopfield Hamiltonian with $M := M(N)$ patterns

$$H_N(\sigma) = -\sum_{i,j=1}^{N} J_{ij}\sigma_i\sigma_j \tag{1}$$

where

$$J_{ij} = \frac{1}{N}\sum_{\mu=1}^{M(N)} \xi_i^\mu \xi_j^\mu$$

and $\sigma_i \in \{-1, 1\}$.
The idea behind the first notion of storage capacity is that a possible retrieval dynamics is a Monte–Carlo dynamics at zero temperature working as follows: Choose a site i at random. Flip the spin σ_i, if flipping lowers the energy (the Hamiltonian) and stay with σ_i otherwise. On a more formal level we define the gradient dynamics T on the energy landscape given by H_N via

$$T : \sigma_i \mapsto sgn(\sum_{j=1}^{N} \sigma_j J_{ij})$$

(where sgn is the sign function) and call a configuration $\sigma = (\sigma_i)_{i\leq N}$ stable if it is a fixed point of T, i.e.

$$\sigma_i = sgn(\sum_{j=1}^{N} \sigma_j J_{ij}) \quad \text{for all } i = 1, \ldots, N$$

which means that σ is a local minimum of the Hamiltonian. The storage capacity in this concept is defined as the greatest number of patterns $M := M(N)$ such that

all the patterns ξ^ν are stable in the above sense (almost surely or with probability converging to one).

The other approach to storage capacity is due to Newman [N88]. It takes into consideration the small errors (mentioned above) we are willing to accept in the restoration of the patterns. So we are satisfied, if the retrieval dynamics converges to a configuration which is not too far away from the original patterns. Thus in this concept a pattern ξ^ν is called stable, if it is close to a local minimum of the Hamiltonian or in other words if it is surrounded by a sufficiently high energy barrier. Technically speaking we will call ξ^ν stable if there exist $\varepsilon > 0$ and $\delta > 0$ such that

$$\inf_{\sigma \in S_\delta(\xi^\nu)} H_N(\sigma) \geq H_N(\xi^\nu) + \varepsilon N. \tag{2}$$

Here the set $S_\delta(\xi^\nu)$ the infimum is taken over is the Hamming sphere of radius δN centered in ξ^ν. Again we will use the notion of storage capacity for the maximal number $M(N)$ of patterns such that (2) holds true for all ξ^ν almost surely.

2 Results in the Case of Unbiased I.I.D. Patterns

In this section we will review the results in the case of unbiased i.i.d. patterns. Most of them go back already to the papers of McEliece et al. [MPRV87] and Newman [N88] and are well-known nowadays. So we will only briefly indicate the basic ideas of the proofs here and refer the interested reader to the original papers or the review article by Petritis [P95] for more detailed informations.
With the definitions introduced above the following results can be proved in the case that the ξ_i^μ are i.i.d. and $P(\xi_i^\mu = 1) = \frac{1}{2}$ (and until otherwise stated we will assume that the patterns are unbiased and i.i.d.).

Theorem 1 *Assume that* $M(N) = \frac{N}{\gamma \log N}$.
Then the following assertions hold true:

1. *If* $\gamma > 6$
$$P(\liminf_{N\to\infty}(\cap_{\nu=1}^{M(N)} T\xi^\nu = \xi^\nu)) = 1$$
i.e. the patterns are almost surely stable.

2. *If* $\gamma \geq 4$
$$P((\cap_{\nu=1}^{M(N)} T\xi^\nu = \xi^\nu)) = 1 - R_N$$
with $\lim_{N\to\infty} R_N = 0$.

3. *If* $\gamma > 2$ *for every fixed* $\nu = 1, \ldots, M$
$$P(T\xi^\nu = \xi^\nu) = 1 - R_N$$
with $\lim_{N\to\infty} R_N = 0$.

Part one of the theorem is contained e.g. in [P95]. Part two of this result first was stated in [MPRV87] and proved in [M92]. Part three has already been proved in [MPRV87].
The idea of the proof is fairly simple. It mainly consists of the observation that according to the definition of the dynamics T the pattern ξ^ν is stable if and only if

$$\sum_{j=1}^{N} \sum_{\mu=1}^{M(N)} \xi_i^\nu \xi_j^\nu \xi_i^\mu \xi_j^\mu \geq 0$$

for all $i = 1, \ldots, N$ (with the convention $sgn(0) = 1$), an application of the exponential Chebyshev–Markov inequality, a computation of the moment generating function

$$E\left(\exp(-t(\sum_{j=1}^{N} \sum_{\substack{\mu=1 \\ \mu \neq \nu}}^{M(N)} \xi_1^\nu \xi_j^\nu \xi_1^\mu \xi_j^\mu))\right) = \cosh(t)^{N\,M(N)} \leq \exp(\frac{1}{2} t^2 N\, M(N))$$

(by the independence of the ξ_i^μ) and a final application of the Borel–Cantelli Lemma. We will give a more explicit proof of a more general statement when proving Theorem 5.

Theorem 1 in other words states that the patterns are fixed points of the gradient dynamics and hence are recognized if one starts with them. But just recalling patterns if they are presented without errors can hardly be called an associative memory. What we would like to have is that even if a pattern is corrupted by a certain percentage of noise the gradient dynamics is able to retrieve this pattern. The following theorem shows that also noised patterns can be successfully reconstructed.

Theorem 2 *(see [KP88],[P95]) Let $r \in [0, \frac{1}{2})$ and for each $\nu = 1, \ldots, M(N)$ let $\tilde{\xi^\nu}$ be an element of the Hamming sphere of radius rN centered at ξ^ν. Assume that $M(N) = (1-2r)^2 \frac{N}{\gamma \log N}$.*
Then:

1. *If $\gamma > 6$*
$$P(\liminf_{N\to\infty}(\cap_{\nu=1}^{M(N)} T\tilde{\xi^\nu} = \xi^\nu)) = 1$$
i.e. the noised patterns are almost surely attracted.

2. *If $\gamma \geq 4$*
$$P((\cap_{\nu=1}^{M(N)} T\tilde{\xi^\nu} = \xi^\nu)) = 1 - R_N$$
with $\lim_{N\to\infty} R_N = 0$.

3. *If $\gamma > 2$ for every fixed $\nu = 1, \ldots, M$*
$$P(T\tilde{\xi^\nu} = \xi^\nu) = 1 - R_N$$
with $\lim_{N\to\infty} R_N = 0$.

The proof of this Theorem follows the same steps as the proof of Theorem 1.

Observe that Theorem 2 basically deals with the case of the so-called "direct convergence" error-correcting power of the Hopfield model, i.e. the convergence to the stored patterns in one iteration. Much more interesting (and technically more involved) is, of course, the question of non-direct convergence, i.e. the number of patterns that can be stored such that the retrieval dynamics starting in a noised pattern eventually converges to the corresponding stored pattern. Already the results in [MPRV87] motivated the authors to conjecture a storage capacity of $\frac{N}{\gamma \log N}$ with again $\gamma = 2, 4$ or 6 depending on whether we concentrate on storing a fixed pattern or all patterns and whether we want convergence in probability or almost surely. This conjecture actually could be proved by [Bu94].

Let us now turn to the second notion of storage capacity. We will see, that if small errors are tolerated, the Hopfield model indeed can store a number of patterns M proportional to the number of neurons N – in agreement with the non-rigorous results of Hopfield [Ho82] and Amit et al. [AGS87] (although the critical α_c is somewhat smaller than what could be expected from the numerical analysis and different concepts of storage capacity are used).

Theorem 3 *There exists an $\alpha_c > 0$ such that if $M(N) \leq \alpha_c N$, then there are $\varepsilon > 0$ and $0 < \delta < 1/2$ such that*

$$P\left(\liminf_{N\to\infty}(\cap_{\nu=1}^{M(N)} \cap_{\sigma\in S_\delta(\xi^\nu)} (H_N(\sigma) \geq H_N(\xi^\nu) + \varepsilon N))\right) = 1$$

where $S_\delta(\xi^\nu)$ is the Hamming sphere of radius δN centered in ξ^ν.

The first proof of this theorem can be found in [N88]. Refined estimates have been obtained in [Lou94] and [T96]. The basic idea is to compute the energy differences between the energy of a fixed pattern ξ^ν and some element in $S_\delta(\xi^\nu)$, to use the exponential Chebyshev–Markov inequality and to replace the variables in the moment generating function by independent $\mathcal{N}(0,1)$- Gaussian random variables. The value of the critical α obtained by this theorem has increased from $\alpha_c = 0.056$ (Newman, [N88]), over $\alpha_c = 0.071$ (Loukianova, [Lou94]) to $\alpha_c = 0.08$ recently proved by Talagrand ([T96]). Again we will see how these ideas are realized in a more explicit proof of a more general statement at the end of this article.

3 The Storage Capacity of The Hopfield Model with Semantically Correlated Patterns

In this section we are going to drop the independence assumption of the previous section. Basically there are two reasonable ways to introduce correlations between the patterns.
One is to consider spatially correlated patterns, i.e. to consider a correlation between ξ_i^ν and ξ_j^ν even if $i \neq j$, but to leave the ξ_i^ν and ξ_j^μ independent for $\mu \neq \nu$. This model may be of interest when storing e.g. images that can be considered to come from a Markov random field. The other type of dependency one may assume is semantical

or sequential dependency among the patterns. That means that we consider random variables ξ_i^ν such that ξ_i^ν and ξ_j^μ still are independent if $i \neq j$, but that we may have correlations between ξ_i^ν and ξ_i^μ even if $\mu \neq \nu$. Such sequences may be interesting if non deterministic sequences of patterns have to be learned, and in some sense every type of human behavior is such a sequence.

Here we will concentrate on the case of semantically correlated patterns as in [Lö96a]. More precisely we assume that the correlation comes from a homogeneous Markov chain and that the patterns ξ_i^μ are correlated in μ but still are independent in i. Such a result is, of course, interesting in its own right, since most realistic situations do not produce independent information. Moreover, one may regard results concerning the Hopfield model with correlated patterns as a step towards showing the universality of the Hopfield model.

So let us assume that the $(\xi_i^\nu)_{i\in\mathbb{N},\nu\in\mathbb{N}}$ form a Markov chain with initial distribution

$$P(\xi_i^1 = x_i^1,\ i = 1,\dots,N) = 2^{-N} \quad \text{for all } x_i^1 \in \{-1,1\} \text{ and all } i = 1,\dots,N. \tag{3}$$

and transition probabilities

$$\begin{aligned} & P(\xi_i^\nu = x_i^\nu | \xi_j^\mu = x_j^\mu, j = 1,\dots,N, \mu = 1,\dots,\nu-1) \\ = \ & P(\xi_i^\nu = x_i^\nu | \xi_i^{\nu-1} = x_i^{\nu-1}) = Q(x_i^{\nu-1}, x_i^\nu). \end{aligned} \tag{4}$$

Here Q denotes a symmetric 2×2 matrix with entries

$$Q = \begin{pmatrix} p & 1-p \\ 1-p & p \end{pmatrix}$$

where $0 < p < 1$ (note that $p = \frac{1}{2}$ is the case of independent patterns).
With this definition our first result concerning correlated patterns reads as follows:

Theorem 4 *Assume the random patterns ξ^ν fulfill (3) and (4) and $M(N) = \frac{N}{\gamma \log N}$. Then for the following assertions hold true:*

1. *If $\gamma > \frac{3(p^2+(1-p)^2)}{p(1-p)}$*
$$P(\liminf_{N\to\infty}(\cap_{\nu=1}^{M(N)} T\xi^\nu = \xi^\nu)) = 1$$
i.e. the patterns are almost surely stable.

2. *If $\gamma \geq \frac{2(p^2+(1-p)^2)}{p(1-p)}$*
$$P((\cap_{\nu=1}^{M(N)} T\xi^\nu = \xi^\nu)) = 1 - R_N$$
with $\lim_{N\to\infty} R_N = 0$.

3. *If $\gamma > \frac{p^2+(1-p)^2}{p(1-p)}$ for every fixed $\nu = 1,\dots,M(N)$*
$$P(T\xi^\nu = \xi^\nu) = 1 - R_N$$
with $\lim_{N\to\infty} R_N = 0$.

Let us only sketch the proof here. For a complete proof we refer the reader to [Lö96]:

Sketch of the Proof: Fix $1 \leq \nu \leq M(N)$. As has been mentioned above the pattern ξ^ν is stable if and only if

$$\sum_{j=1}^{N} \sum_{\mu=1}^{M(N)} \xi_i^\nu \xi_j^\nu \xi_i^\mu \xi_j^\mu \geq 0$$

for all $i = 1, \ldots, N$.
Hence for by the identical distribution of the ξ_i^μ for different i and the exponential Chebyshev-inequality we obtain all $t \geq 0$

$$\begin{aligned} P(\xi^\nu \text{ is not stable }) &\leq N P(\sum_{j=1}^{N} \sum_{\substack{\mu=1 \\ \mu \neq \nu}}^{M(N)} \xi_1^\nu \xi_j^\nu \xi_1^\mu \xi_j^\mu \leq -N) \\ &= N e^{-tN} \left(E \left(\exp(-t(\sum_{\substack{\mu=1 \\ \mu \neq \nu}}^{M(N)} \xi_1^\nu \xi_2^\nu \xi_1^\mu \xi_2^\mu)) \right) \right)^N \end{aligned} \tag{5}$$

Now putting $Y_\mu := \xi_1^\mu \xi_2^\mu$ and calculating the expectation in (5) leads to

$$\begin{aligned} & E(\exp(-t \sum_{\substack{\mu=1, \\ \mu \neq \nu}}^{M(N)} Y_\mu Y_\nu)) \\ = & \sum_{\substack{y_1=-1,1, \\ y_M=-1,1}} \Pi_L^{\nu-1}(y_1, 1) \Pi_R^{M-\nu}(1, y_M) \\ = & \left(\begin{pmatrix} 1 \\ 1 \end{pmatrix} \Pi_L^{\nu-1} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right) \times \left(\begin{pmatrix} 1 \\ 0 \end{pmatrix} \Pi_R^{M-\nu}(1, \cdot) \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right) \leq \lambda_1^{M-1} \end{aligned}$$

where

$$\Pi_L := \begin{pmatrix} qe^{-t} & (1-q)e^{-t} \\ (1-q)e^{t} & qe^{t} \end{pmatrix},$$

$(\Pi_L)^t = \Pi_R$, and λ_1 is the largest eigenvalues of Π_L. Observe that

$$\lambda_1 = q \cosh(t) + \sqrt{1 - 2q + q^2 \cosh^2(t)} \tag{6}$$

Hence we arrive at

$$P(\xi^\nu \text{ is not stable }) \leq N e^{-tN} \lambda_1^{(M(N)-1)N}.$$

Moreover, expanding the root in (6) using $\sqrt{(1+x)} \leq 1 + \frac{x}{2}$ and approximating the hyperbolic functions contained in (6) by their leading two terms yields

$$\lambda_1 \leq 1 + t^2 \frac{q}{2(1-q)} + \mathcal{O}(t^4) \leq \exp(t^2 \frac{q}{2(1-q)})(1 + \mathcal{O}(t^4)).$$

Choosing $t = \frac{1-q}{qM(N)}$ gives

$$P(\xi^\nu \text{ is not stable }) \leq N \exp(-\frac{1-q}{2q}\frac{N}{M(N)})(1+\mathcal{O}(t^4))^{M(N)N}.$$

So if $M(N) = \frac{N}{\gamma \log N}$ the last factor on the right hand side can be bounded by $\exp(const.\frac{(\log N)^4}{N^2})$ which is converging to one. Hence the right hand side of the inequality is bounded by $const.\, N^{1-\frac{(1-q)\gamma}{2q}}$ which, for $\gamma > \frac{2q}{1-q} = \frac{p^2+(1-p)^2}{p(1-p)}$, converges to zero and therefore yields part three of the theorem.

For the other two parts observe that the bounds obtained above do not depend on ν. Thus

$$P(\exists \nu : \xi^\nu \text{ is not stable }) \leq M(N)N \exp(-\frac{1-q}{2q}\frac{N}{M(N)})\mathcal{O}(1)$$

So putting again $M(N) = \frac{N}{\gamma \log N}$ this time with $\gamma > \frac{6q}{1-q} = \frac{3(p^2+(1-p)^2)}{p(1-p)}$ leads to the converging series $\sum \frac{1}{N^\kappa \log N}$ for an $\kappa > 1$ and thus proves part one of the theorem by the Borel–Cantelli Lemma. The choice of $\gamma \geq \frac{4q}{1-q} = \frac{2(p^2+(1-p)^2)}{p(1-p)}$ yields

$$P(\exists \nu : \xi^\nu \text{ is not stable }) \to 0$$

and therefore part two of the theorem. □

Observe that the bounds obtained in Theorem 4 are decreasing functions of the correlation. This in a way reflects the idea that the basic reason why the Hopfield model works well as an associative memory in the case of i.i.d. patterns is that such patterns tend to be "nearly orthogonal" which more precisely means that the overlap $\frac{1}{N}\sum_{i=1}^N \xi_i^\mu \xi_i^\nu$ for $\mu \neq \nu$ if of order $N^{-\frac{1}{2}}$ (and it is e.g. quickly checked that the Hopfield model indeed can store N orthogonal patterns). For sequences of correlated patterns such a behavior cannot be expected. However, since Markov chains have exponentially decreasing correlation the dependencies do not influence the storage capacities too heavily in our case.

Let us also mention that there is, of course, a version of Theorem 2 for the case of patterns fulfilling (3) and (4). The value of γ there is the one which could be expected from Theorems 2 and 4 (also see [Lö96a]).

With the second notion of storage capacity we obtain the following result

Theorem 5 *Suppose that the random patterns fulfill (3) and (4). There exists an $\alpha_c > 0$ (depending on p) such that if $M(N) \leq \alpha_c N$, then there are $\varepsilon > 0$ and $0 < \delta < 1/2$ such that*

$$P\left(\liminf_{N\to\infty}(\cap_{\nu=1}^{M(N)} \cap_{\sigma\in S_\delta(\xi^\nu)} (H_N(\sigma) \geq H_N(\xi^\nu) + \varepsilon N))\right) = 1$$

where $S_\delta(\xi^\nu)$ is the Hamming sphere of radius δN centered in ξ^ν.

We present the proof as given in [Lö96a].

Proof:
The main steps of the proof consist of a centering of the patterns and by replacing them by appropriate Gaussian random variables. Although this basic idea is fairly standard in the context of storage capacity estimates (see e.g. [N88], [BG92]) in our situation the computations become technically quite involved.

We set

$$h_N(\sigma,\delta) := \inf_{\sigma' \in S_\delta(\sigma)} H_N(\sigma').$$

First of all observe that

$$\begin{aligned} & P\left(\{\cap_{\nu=1}^{M(N)}(h_N(\xi^\nu,\delta) \geq H_N(\xi^\nu)+\varepsilon N)\}^c\right) \\ \leq & \sum_{J:|J|=\delta N} \sum_{\nu=1}^{M(N)} P\left(H_N(\xi_J^\nu) - H_N(\xi^\nu) \leq \varepsilon N\right) \end{aligned}$$

where ξ_J^ν denotes a configuration differing from ξ^ν exactly in the coordinates J and δ is chosen in such a way that δN is an integer.
Let us keep ν fixed in the sequel and note that

$$H_N(\xi_J^\nu) - H_N(\xi^\nu) = \frac{2}{N} \sum_{\mu \neq \nu} \sum_{i \in J, j \notin J} \xi_i^\nu \xi_j^\nu \xi_i^\mu \xi_j^\mu + 2\delta(1-\delta).$$

Thus by the exponential Chebyshev-Markov inequality for any $t \geq 0$

$$P\left(H_N(\xi_J^\nu) - H_N(\xi^\nu) \leq \varepsilon N\right) \leq e^{-t\varepsilon' N} E\left(\exp\left(-\frac{t}{N} \sum_{\mu \neq \nu} \sum_{i \in J, j \notin J} \xi_i^\nu \xi_j^\nu \xi_i^\mu \xi_j^\mu\right)\right)$$

where we have set $\varepsilon' = -\varepsilon/2 + \delta(1-\delta)$.
Let us moreover assume that $\xi_i^\nu = 1$ for all $i = 1, \ldots, N$ (this can be done without loss of generality since the initial situation is completely symmetric). Then the sum in the exponent of the moment generating function can be split into two parts:

$$\sum_{\mu \neq \nu} \sum_{i \in J, j \notin J} \xi_i^\mu \xi_j^\mu = \sum_{\mu > \nu} \sum_{i \in J, j \notin J} \xi_i^\mu \xi_j^\mu + \sum_{\mu < \nu} \sum_{i \in J, j \notin J} \xi_i^\mu \xi_j^\mu \tag{7}$$

which, conditioned on $\xi_i^\nu = 1$ for all $i = 1, \ldots, N$, are independent. Introducing

$$\overline{\xi_i^\mu} = \xi_i^\mu - (2p-1)\xi_i^{\mu-1}. \tag{8}$$

we can express the first sum on the right hand side of (7) as

$$\begin{aligned} & \sum_{i \in J, j \notin J} \sum_{\mu > \nu}^{M} \xi_i^\mu \xi_j^\mu \\ = & \sum_{i \in J, j \notin J} \left(\sum_{\mu_1, \mu_2 > \nu}^{M} a_{\mu_1,\mu_2} \overline{\xi_i^{\mu_1}}\, \overline{\xi_j^{\mu_2}} + \sum_{\mu > \nu}^{M} a_{\mu,\nu}(\overline{\xi_i^\mu} + \overline{\xi_j^\mu}) + \sum_{n=0}^{M-\nu-1} (2p-1)^{2n} \right), \end{aligned}$$

where

$$a_{\mu_1,\mu_2} := \sum_{n=0}^{M-\max\{\mu_1,\mu_2\}} (2p-1)^{2n+|\mu_1-\mu_2|} \tag{9}$$

for $\mu_1, \mu_2 \geq \nu$, $(\mu_1, \mu_2) \neq (\nu, \nu)$. Note that $a_{\mu_1,\mu_2} = a_{\mu_2,\mu_1}$.

For the second sum in (7) we observe that reversing the chains $(\xi_i^\mu)_{\mu<\nu}$ $(i = 1, \ldots, N)$ does not change their distribution. So applying the same transformation as above to the reversed Markov chains $(\xi_i^\mu)_{\mu<\nu}$ $(i = 1, \ldots, N)$ yields

$$\begin{aligned} E(\exp(-\frac{t}{N}\sum_{i\in J, j\notin J}\sum_{\mu\neq\nu}\xi_i^\mu\xi_j^\mu)) &= \exp(-\frac{t}{N}\sum_{i\in J,j\notin J}(\sum_{n=0}^{M-\nu-1}(2p-1)^{2n} + \sum_{n=0}^{\nu-1}(2p-1)^{2n})) \\ \times \quad E\left(\exp\left(-\frac{t}{N}\sum_{i\in J,j\notin J}\left(\sum_{\mu>\nu}a_{\mu,\nu}(\overline{\xi_i^\mu}+\overline{\xi_j^\mu}) + \sum_{\mu<\nu}\tilde{a}_{\mu,\nu}(\overline{\xi_i^\mu}+\overline{\xi_j^\mu}) + \right.\right.\right. \\ \left.\left.\left. + \sum_{\mu_1,\mu_2>\nu}a_{\mu_1,\mu_2}\overline{\xi_i^{\mu_1}}\,\overline{\xi_j^{\mu_2}} + \sum_{\mu_1,\mu_2<\nu}\tilde{a}_{\mu_1,\mu_2}\overline{\xi_i^{\mu_1}}\,\overline{\xi_j^{\mu_2}}\right)\right)\right), \end{aligned}$$

where

$$\tilde{a}_{\mu_1,\mu_2} := \sum_{n=0}^{\nu-1-\min\{\nu-\mu_1,\nu-\mu_2\}} (2p-1)^{2n+|\mu_1-\mu_2|}. \tag{10}$$

Using the independence of the initial part and the tail part of the Markov chains mentioned above together with Hölder's inequality to split up the moment generating function of the linear part from the moment generating function of the genuine quadratic form we obtain for all $\lambda > 1$

$$\begin{aligned} & E\left(\exp\left(-\frac{t}{N}\sum_{\mu\neq\nu}\sum_{i\in J,j\notin J}\xi_i^\nu\xi_j^\nu\xi_i^\mu\xi_j^\mu\right)\right) \\ \leq \quad & \exp(-tN\delta(1-\delta)(\sum_{n=0}^{M-\nu-1}(2p-1)^{2n} + \sum_{n=0}^{\nu-1}(2p-1)^{2n})) \qquad (11) \\ \times \quad & \left(E\left(\exp\left(-\frac{t}{N}\frac{\lambda}{\lambda-1}\sum_{\mu>\nu}a_{\mu,\nu}\sum_{i\in J,j\notin J}(\overline{\xi_i^\mu}+\overline{\xi_j^\mu})\right)\right)\right)^{\frac{\lambda-1}{\lambda}} \times \\ \times \quad & \left(E\left(\exp\left(-\frac{t}{N}\frac{\lambda}{\lambda-1}\sum_{\mu<\nu}\tilde{a}_{\mu,\nu}\sum_{i\in J,j\notin J}(\overline{\xi_i^\mu}+\overline{\xi_j^\mu})\right)\right)\right)^{\frac{\lambda-1}{\lambda}} \\ \times \quad & \left(E\left(\exp\left(-\frac{t}{N}\lambda\sum_{i\in J,j\notin J}\sum_{\mu_1,\mu_2>\nu}a_{\mu_1,\mu_2}\overline{\xi_i^{\mu_1}}\,\overline{\xi_j^{\mu_2}}\right)\right)\right)^{\frac{1}{\lambda}} \end{aligned}$$

$$\times \quad \left(E\left(\exp\left(-\frac{t}{N}\lambda \sum_{i\in J, j\notin J} \sum_{\mu_1,\mu_2<\nu} \tilde{a}_{\mu_1,\mu_2} \overline{\xi_i^{\mu_1}}\,\overline{\xi_j^{\mu_2}} \right)\right)\right)^{\frac{1}{\lambda}} .$$

We now have to estimate the factors on the right hand side of (11). Note that for M large enough

$$(\sum_{n=0}^{M-\nu-1} (2p-1)^{2n} + \sum_{n=0}^{\nu-1}(2p-1)^{2n})) \geq \frac{1}{C'(1-(2p-1)^2)}.$$

for any $C' > 1$

To treat the other terms let us agree on the following notation: With $E_I^{I'}$ (where $I \subset \{1,\ldots,N\}$ and $I' \subset \{1,\ldots,M\}$) we denote the integration with respect to those random variables ξ_i^μ with $i \in I$ and $\mu \in I'$. Especially, if we drop the upper or lower indices we will usually mean the expectation with respect to all the random variables occuring in the argument of the integral. By the independence of the coordinate processes and the identical distribution of the ξ_i^μ we obtain for the moment generating function of the linear part

$$\begin{aligned} & E\left(\exp\left(-\frac{t}{N}\frac{\lambda}{\lambda-1} \sum_{\mu>\nu} a_{\mu,\nu} \sum_{i\in J, j\notin J} (\overline{\xi_i^\mu} + \overline{\xi_j^\mu})\right)\right) \\ = \quad & \left[E\left(\exp(-\frac{t}{N}\frac{\lambda}{\lambda-1} \sum_{\mu>\nu} a_{\mu,\nu}\overline{\xi_1^\mu})\right)\right]^{\delta(1-\delta)N^2} \end{aligned}$$

The expectation above can now be estimated as follows

$$\begin{aligned} & E\left(\exp\left(-\frac{t}{N}\frac{\lambda}{\lambda-1} \sum_{\mu>\nu} a_{\mu,\nu}\overline{\xi_1^\mu}\right)\right) \\ = \quad & E^{\nu<\mu\leq M-1}\left(\exp\left(-\frac{t}{N}\frac{\lambda}{\lambda-1} \sum_{\mu>\nu}^{M-1} a_{\mu,\nu}\overline{\xi_1^\mu}\right)\right) E^M\left(\exp(-\frac{t}{N}\frac{\lambda}{\lambda-1} a_{M,\nu}\overline{\xi_1^M})\right) \\ = \quad & E^{\nu<\mu\leq M-1}\left(\exp\left(-\frac{t}{N}\frac{\lambda}{\lambda-1} \sum_{\mu>\nu}^{M-1} a_{\mu,\nu}\overline{\xi_1^\mu}\right)\right) \times \\ \times \quad & \left(p\exp(-2\frac{t}{N}\frac{\lambda}{\lambda-1} a_{M,\nu}(1-p)\xi_1^{M-1}) + (1-p)\exp(2\frac{t}{N}\frac{\lambda}{\lambda-1} a_{M,\nu}p\xi_1^{M-1}\right) \\ \leq \quad & E^{\nu<\mu\leq M-1}\left(\exp\left(-\frac{t}{N}\frac{\lambda}{\lambda-1} \sum_{\mu>\nu}^{M-1} a_{\mu,\nu}\overline{\xi_1^\mu}\right)\right) \cosh(\frac{t}{N}\frac{\lambda}{\lambda-1} a_{M,\nu}(1+|2p-1|)\xi_1^{M-1}) \\ \leq \quad & E^{\nu<\mu\leq M-1}\left(\exp\left(-\frac{t}{N}\frac{\lambda}{\lambda-1} \sum_{\mu>\nu}^{M-1} a_{\mu,\nu}\overline{\xi_1^\mu}\right)\right) \exp(\frac{1}{2}\frac{t^2}{N^2}(\frac{\lambda}{\lambda-1})^2 a_{M,\nu}^2(1+|2p-1|)^2) \end{aligned}$$

where we have used $|\xi_1^{M-1}| = 1$,

$$p\exp(-2(1-p)t) + (1-p)\exp(2pt) \leq \cosh((1+|2p-1|)t)$$

for all $0 < p < 1$ and all $t \in \mathbb{R}$ and finally

$$\cosh(x) \leq \exp(x^2/2).$$

Integrating the other variables in the same way gives

$$\begin{aligned} & E\left(\exp\left(-\frac{t}{N}\frac{\lambda}{\lambda-1}\sum_{\mu>\nu} a_{\mu,\nu}\overline{\xi_1^\mu}\right)\right) \\ \leq\ & \exp(\frac{1}{2}\frac{t^2}{N^2}(\frac{\lambda}{\lambda-1})^2(1+|2p-1|)^2\sum_{\mu>\nu}^{M} a_{\mu,\nu}^2) \\ \leq\ & \exp\left(\frac{1}{2}\frac{t^2}{N^2}(\frac{\lambda}{\lambda-1})^2(1+|2p-1|)^2\frac{1}{(1-(2p-1)^2)^3}\right). \end{aligned}$$

So altogether we arrive at

$$\begin{aligned} & E\left(\exp\left(-\frac{t}{N}\frac{\lambda}{\lambda-1}\sum_{\mu>\nu} a_{\mu,\nu}\sum_{i\in J, j\notin J}(\overline{\xi_i^\mu}+\overline{\xi_j^\mu})\right)\right) \\ \leq\ & \exp\left(\frac{1}{2}t^2\delta(1-\delta)(\frac{\lambda}{\lambda-1})^2(1+|2p-1|)^2\frac{1}{(1-(2p-1)^2)^3}\right) \end{aligned}$$

Thus applying the same techniques to the second linear term on the right hand side of (11) we obtain

$$\begin{aligned} & \left(E\left(\exp\left(-\frac{t}{N}\frac{\lambda}{\lambda-1}\sum_{\mu>\nu} a_{\mu,\nu}\sum_{i\in J, j\notin J}(\overline{\xi_i^\mu}+\overline{\xi_j^\mu})\right)\right)\right)^{\frac{\lambda-1}{\lambda}} \times \\ \times\ & \left(E\left(\exp\left(-\frac{t}{N}\frac{\lambda}{\lambda-1}\sum_{\mu<\nu} \tilde{a}_{\mu,\nu}\sum_{i\in J, j\notin J}(\overline{\xi_i^\mu}+\overline{\xi_j^\mu})\right)\right)\right)^{\frac{\lambda-1}{\lambda}} \\ \leq\ & \exp\left(t^2\delta(1-\delta)(\frac{\lambda}{\lambda-1})(1+|2p-1|)^2\frac{1}{(1-(2p-1)^2)^3}\right) \end{aligned}$$

We will see that due to our final choice of t this factor will have a negligible contribution to the final estimate (which might have been expected by just counting the number of linear terms and comparing it to the number of terms in the genuine quadratic form.

The moment generating function of the quadratic form is treated similarly using the independence of the ξ_i^μ for different i to replace them by Gaussian random variables:

$$E\left(\exp\left(-\frac{t}{N}\lambda\sum_{i\in J, j\notin J}\sum_{\mu_1,\mu_2>\nu} a_{\mu_1,\mu_2}\overline{\xi_i^{\mu_1}}\,\overline{\xi_j^{\mu_2}}\right)\right)$$

$$
\begin{aligned}
&= E^{\nu<\mu_1,\mu_2\leq M-1}E^M_{J^c}\left(\exp\left(-\frac{t}{N}\lambda\sum_{i\in J,j\notin J}\sum_{\mu_2>\nu}^{M}\sum_{\mu_1>\nu}^{M-1}a_{\mu_1,\mu_2}\overline{\xi_i^{\mu_1}}\,\overline{\xi_j^{\mu_2}}\right)\right.\\
&\quad \left.E^M_J\left(\exp(-\frac{t}{N}\lambda\sum_{i\in J,j\notin J}\overline{\xi_i^M}\sum_{\mu_2=\nu+1}^{M}a_{M,\mu_2}\overline{\xi_j^{\mu_2}})\right)\right)\\
&= E^{\nu<\mu_1,\mu_2\leq M-1}E^M_{J^c}\left(\exp\left(-\frac{t}{N}\lambda\sum_{i\in J,j\notin J}\sum_{\mu_2>\nu}^{M}\sum_{\mu_1>\nu}^{M-1}a_{\mu_1,\mu_2}\overline{\xi_i^{\mu_1}}\,\overline{\xi_j^{\mu_2}}\right)\right.\\
&\quad \left.\prod_{i\in J}E^M_{\{i\}}\left(\exp(-\frac{t}{N}\lambda\overline{\xi_i^M}\sum_{j\notin J}\sum_{\mu_2=\nu+1}^{M}a_{M,\mu_2}\overline{\xi_j^{\mu_2}})\right)\right)\\
&\leq E^{\nu<\mu_1,\mu_2\leq M-1}E^M_{J^c}\left(\exp\left(-\frac{t}{N}\lambda\sum_{i\in J,j\notin J}\sum_{\mu_2>\nu}^{M}\sum_{\mu_1>\nu}^{M-1}a_{\mu_1,\mu_2}\overline{\xi_i^{\mu_1}}\,\overline{\xi_j^{\mu_2}}\right)\right)\times\\
&\times \prod_{i\in J}\exp\left(\frac{1}{2}\frac{t^2}{N^2}\lambda^2(1+|2p-1|)^2(\sum_{j\notin J}\sum_{\mu_2=\nu+1}^{M}a_{M,\mu_2}\overline{\xi_j^{\mu_2}})^2\right)\\
&= E^{\nu<\mu_1,\mu_2\leq M-1}E^M_{J^c}\left(\exp\left(-\frac{t}{N}\lambda\sum_{i\in J,j\notin J}\sum_{\mu_2>\nu}^{M}\sum_{\mu_1>\nu}^{M-1}a_{\mu_1,\mu_2}\overline{\xi_i^{\mu_1}}\,\overline{\xi_j^{\mu_2}}\right)\right)\times\\
&\times \prod_{i\in J}E_{z_i^M}\exp\left(z_i^M\frac{t}{N}\lambda(1+|2p-1|)\sum_{i\in J,j\notin J}\sum_{\mu_2=\nu+1}^{M}a_{M,\mu_2}\overline{\xi_j^{\mu_2}}\right)\\
&= E^{\nu<\mu_1,\mu_2\leq M-1}E^M_{J^c}\left(\exp\left(-\frac{t}{N}\lambda\sum_{i\in J,j\notin J}\sum_{\mu_2>\nu}^{M}\sum_{\mu_1>\nu}^{M-1}a_{\mu_1,\mu_2}\overline{\xi_i^{\mu_1}}\,\overline{\xi_j^{\mu_2}}\right)\right)\times\\
&\times E_{z_J^M}\exp\left(\frac{t}{N}\lambda(1+|2p-1|)\sum_{i\in J,j\notin J}\sum_{\mu_2=\nu+1}^{M}a_{M,\mu_2}z_i^M\overline{\xi_j^{\mu_2}}\right)
\end{aligned}
$$

where z_i^M are Gaussian random variables with expectation 0 and identity covariance matrix independent of the ξ_i^μ, $E_{z_i^M}$ denotes the expectation with respect to z_i^M, and finally $E_{z_J^M}$ denotes the expectation with respect to the vector $(z_i^M)_{i\in J}$. Here we have used the well known identity

$$\exp(\frac{1}{2}x^2)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp(xy-\frac{1}{2}y^2)dy.$$

Interchanging the order of integration and using the above technique on every ξ_i^μ we are now able to consecutively replace all the variables ξ_i^μ by Gaussian random variables z_i^μ with expectation zero and identity covariance matrix. This leads to

$$E\left(\exp\left(-\frac{t}{N}\lambda\sum_{i\in J,j\notin J}\sum_{\mu_1,\mu_2>\nu}a_{\mu_1,\mu_2}\overline{\xi_i^{\mu_1}}\,\overline{\xi_j^{\mu_2}}\right)\right)$$

$$
\begin{aligned}
&\leq E_z\left(\exp\left(\frac{t}{N}\lambda(1+|2p-1|)^2\sum_{i\in J, j\notin J}\sum_{\mu_1,\mu_2>\nu} a_{\mu_1,\mu_2} z_i^{\mu_1} z_j^{\mu_2}\right)\right)\\
&\leq E_z\left(\exp\left(t\lambda(1+|2p-1|)^2\sqrt{\delta(1-\delta)}\sum_{\mu_1,\mu_2>\nu} a_{\mu_1,\mu_2} z^{\mu_1}\overline{z^{\mu_2}}\right)\right)\\
&= E_z\left(\exp\frac{1}{2}\left(t\lambda(1+|2p-1|)^2\sqrt{\delta(1-\delta)}\langle z,\hat{A}z\rangle\right)\right)
\end{aligned}
$$

where (by normalizing) $(z^\mu)_{\mu=\nu+1,\ldots,M}$ and $(\overline{z^\mu})_{\mu=\nu+1,\ldots,M}$ are now Gaussian random variables with expectation 0 and identity covariance matrix, z denotes the vector of the $(z^\mu, \overline{z^\mu})$ and E_z is integration with respect to z. Finally $\hat{A}$ is an $2(M-\nu)\times 2(M-\nu)$-matrix with entries

$$
\hat{A} = \left(\begin{array}{c|c} 0 & A \\ \hline A & 0 \end{array}\right)
$$

and the $(M-\nu)\times(M-\nu)$-matrix A is given by

$$
A = (A_{\mu_1,\mu_2}) = (a_{\mu_1-\nu,\mu_2-\nu}).
$$

.

Observe that the above integral only exists if t is small enough (i.e. if $Id - t\lambda(1+|2p-1|)^2\sqrt{\delta(1-\delta)}\hat{A}$ is positive definite) and in this case it equals the inverse of the square-root of the determinant of $Id - t\lambda(1+|2p-1|)^2\sqrt{\delta(1-\delta)}\hat{A}$. On the other hand this determinant can be estimated since trivially the identity matrix commutes with $\hat{A}$. Thus

$$
\begin{aligned}
&\det(Id - t\lambda(1+|2p-1|)^2\sqrt{(\delta(1-\delta)}\hat{A}) = \prod_{i=1}^{2(M-\nu)}\varrho_i\\
&= \prod_{i=1}^{2(M-\nu)}(1 - t\lambda(1+|2p-1|)^2\sqrt{\delta(1-\delta)}\alpha_i)
\end{aligned}
$$

where the ϱ_i are the eigenvalues of $Id - t\lambda(1+|2p-1|)^2\sqrt{\delta(1-\delta)}\hat{A}$ and the α_i are the eigenvalues of $\hat{A}$. Moreover note that $\hat{A}$ has a symmetric spectrum, i.e. if α_i is an eigenvalue of $\hat{A}$ then so is $-\alpha_i$ (which can be seen from the fact that if $v = (v_1,\ldots,v_{M-\nu},v_{M-\nu+1},\ldots v_{2(M-\nu)})$ is an eigenvector for the eigenvalue α_i then $\tilde{v} = (-v_1,\ldots,-v_{M-\nu},v_{M-\nu+1},\ldots v_{2(M-\nu)})$ is an eigenvector for $-\alpha_i$). Therefore

$$
\begin{aligned}
&\det(Id - t\lambda(1+|2p-1|)^2\sqrt{(\delta(1-\delta)}\hat{A}) = \prod_{i=1}^{M-\nu}(1 - t^2\lambda^2(1+|2p-1|)^4\delta(1-\delta)\alpha_i^2)\\
&\geq (1 - t^2\lambda^2(1+|2p-1|)^4\delta(1-\delta)\alpha_{\max}^2)^{M-\nu}
\end{aligned}
$$

where the product is taken over all non-negative eigenvalues and $\alpha_{\max}$ denotes the maximum eigenvalue of $\hat{A}$. This maximum eigenvalue by Gershgorin's theorem can be bounded by the maximum row sum, i.e.

$$
\alpha_{\max} \leq \max_{\mu_1}\sum_{\mu_2}|a_{\mu_1,\mu_2}| \leq \frac{1}{1-(2p-1)^2}\frac{2}{1-|2p-1|}.
$$

Plugging that into our estimates gives

$$\begin{aligned} & E_z\left(\exp\left(\frac{t}{N}\lambda(1+|2p-1|)\sqrt{\delta(1-\delta)}\sum_{\mu_1,\mu_2>\nu}a_{\mu_1,\mu_2}z^{\mu_1}z^{\mu_2}\right)\right) \\ \leq & \left(\frac{1}{\sqrt{1-t^2\lambda^2(1+|2p-1|)^4\delta(1-\delta)(\frac{1}{1-(2p-1)^2}\frac{2}{1-|2p-1|})^2}}\right)^{M-\nu}. \end{aligned}$$

where we have assumed that t is so small that the latter quantity is real.

Thus repeating the estimate for the moment generating function of the second quadratic form and setting $M=\alpha N$

$$\begin{aligned} & P\left(H_N(\xi_J^\nu)-H_N(\xi^\nu)\leq\varepsilon N\right) \\ \leq & \inf_{t^*\geq t\geq 0}\exp\left(-t\varepsilon' N - tN\delta(1-\delta)\frac{1}{C'(1-(2p-1)^2)}\right) \\ \times & \exp\left(-\log\left(1-t^2\lambda^2\frac{\delta(1-\delta)}{(1-(2p-1)^2)^2}\frac{4(1+|2p-1|)^4}{(1-|2p-1|)^2}\right)\frac{M-\nu}{2}\right) \\ \times & \exp\left(-\log\left(1-t^2\lambda^2\frac{\delta(1-\delta)}{(1-(2p-1)^2)^2}\frac{4(1+|2p-1|)^4}{(1-|2p-1|)^2}\right)\frac{\nu}{2}\right) \\ \times & \exp\left(t^2\delta(1-\delta)(\frac{\lambda}{\lambda-1})(1+|2p-1|)^2\frac{1}{(1-(2p-1)^2)^3}\right) \\ = & \inf_{t^*\geq t\geq 0}\exp\left(-t\varepsilon' N - tN\delta(1-\delta)\frac{1}{C'(1-(2p-1)^2)}\right) \\ \times & \exp\left(-\log\left(1-t^2\lambda^2\frac{\delta(1-\delta)}{(1-(2p-1)^2)^2}\frac{4(1+|2p-1|)^4}{(1-|2p-1|)^2}\right)\frac{M}{2}\right) \\ \times & \exp\left(t^2\delta(1-\delta)(\frac{\lambda}{\lambda-1})(1+|2p-1|)^2\frac{1}{(1-(2p-1)^2)^3}\right) \end{aligned}$$

where $t^*=\frac{(1-(2p-1)^2)(1-|2p-1|)}{2\lambda(1+|2p-1|)^2}\sqrt{\frac{1}{\delta(1-\delta)}}$.

Finally by Stirling's formula (to bound the binomial coefficient) and the above estimate

$$\begin{aligned} & \sum_{J:|J|=\delta N}\sum_{\nu=1}^{M(N)}P\left(H_N(\xi_J^\nu)-H_N(\xi^\nu)\leq\varepsilon N\right) \\ \leq & M(N)\binom{N}{\delta N}\exp\left(-t\varepsilon' N - tN\delta(1-\delta)\frac{1}{C'(1-(2p-1)^2)}\right)\times \\ \times & \exp\left(-\log(1-t^2\lambda^2\frac{\delta(1-\delta)}{(1-(2p-1)^2)^2}\frac{4(1+|2p-1|)^4}{(1-|2p-1|)^2})\frac{\alpha}{2}N\right) \\ \times & \exp\left(t^2\delta(1-\delta)(\frac{\lambda}{\lambda-1})(1+|2p-1|)^2\frac{1}{(1-(2p-1)^2)^3}\right) \\ \leq & \alpha N\inf_{t^*\geq t\geq 0}\exp\left((-\delta\log\delta-(1-\delta)\log(1-\delta))N\right)\times \\ \times & \exp\left(-t\varepsilon' N - tN\delta(1-\delta)\frac{1}{C'(1-(2p-1)^2)}\right) \end{aligned}$$

$$\times \quad \exp\left(-\log(1-t^2\lambda^2\frac{\delta(1-\delta)}{(1-(2p-1)^2)^2}\frac{4(1+|2p-1|)^4}{(1-|2p-1|)^2})\frac{\alpha}{2}N\right)$$
$$\times \quad \exp\left(t^2\delta(1-\delta)(\frac{\lambda}{\lambda-1})(1+|2p-1|)^2\frac{1}{(1-(2p-1)^2)^3}\right)$$

and we have to find an admissible t (i.e. $0 \le t \le t^*$) and values of δ and α such that the above exponent becomes negative. To this end first of all note that for all admissible t

$$\exp\left(t^2\delta(1-\delta)(\frac{\lambda}{\lambda-1})(1+|2p-1|)^2\frac{1}{(1-(2p-1)^2)^3}\right) = \mathcal{O}(1)$$

and therefore this term does not influence the convergence (as promised above).

Moreover if $t^2\lambda^2\frac{\delta(1-\delta)}{(1-(2p-1)^2)^2}\frac{4(1+|2p-1|)^4}{(1-|2p-1|)} \le 3/4$

$$\frac{1}{\sqrt{1-t^2\lambda^2\frac{\delta(1-\delta)}{(1-(2p-1)^2)^2}\frac{4(1+|2p-1|)^4}{(1-|2p-1|)^2}}}$$
$$\le \quad \exp\left(4t^2\lambda^2\frac{\delta(1-\delta)}{(1-(2p-1)^2)^2}\frac{(1+|2p-1|)^4}{(1-|2p-1|)^2}\right).$$

and hence up to terms of order one $\sum_{J:|J|=\delta N}\sum_{\nu=1}^{M(N)} P\left(H_N(\xi_J^\nu) - H_N(\xi^\nu) \le \varepsilon N\right)$ can be bounded by

$$\exp\left((-\delta\log\delta-(1-\delta)\log(1-\delta))N - t\varepsilon' N - tN\delta(1-\delta)\frac{1}{C'(1-(2p-1)^2)}\right.$$
$$\left.-\log(1-t^2\lambda^2\frac{\delta(1-\delta)}{(1-(2p-1)^2)^2}\frac{4(1+|2p-1|)^4}{(1-|2p-1|)^2})\frac{\alpha}{2}N\right)$$
$$\le \quad \exp\left((-\delta\log\delta-(1-\delta)\log(1-\delta))N - t\varepsilon' N - tN\delta(1-\delta)\frac{1}{C'(1-(2p-1)^2)}\right.$$
$$\left.+4t^2\lambda^2\frac{\delta(1-\delta)}{(1-(2p-1)^2)^2}\frac{(1+|2p-1|)^4}{(1-|2p-1|)^2}\alpha N\right)$$

if

$$t \le t^{**} := \frac{(1-(2p-1)^2)(1-|2p-1|)}{4\lambda(1+|2p-1|)^2}\sqrt{\frac{3}{\delta(1-\delta)}}.$$

Choosing ε very small the exponent is minimized by a t which is close to

$$t_{\min} = \frac{1}{\alpha}\frac{1}{8\lambda^2(1+|2p-1|)^4}((1-(2p-1)^2)+\frac{1}{C'})(1-(2p-1)^2)(1-|2p-1|)^2.$$

Observe that $t_{\min} \le t^{**}$ if

$$\alpha \ge \sqrt{\delta(1-\delta)}\frac{1}{\sqrt{3}\lambda(1+|2p-1|)^2}(1-(2p-1)^2+\frac{1}{C'})(1-|2p-1|). \qquad (12)$$

On the other hand inserting $t_{\min}$ into the essential part of the exponent and choosing ε sufficiently small gives (for the exponent)

$$
\begin{aligned}
& (-\delta\log\delta-(1-\delta)\log(1-\delta))N - t_{\min}\varepsilon' N - t_{\min}N\delta(1-\delta)\frac{1}{C'(1-(2p-1)^2)} \\
+ \quad & 4t_{\min}^2\lambda^2\frac{\delta(1-\delta)}{1-(2p-1)^2}\frac{(1+|2p-1|)^4}{1-|2p-1|}\alpha N \qquad (13)\\
\leq \quad & (-\delta\log\delta-(1-\delta)\log(1-\delta))N - \gamma\frac{\delta(1-\delta)(1-|2p-1|)^2(1-(2p-1)^2+\frac{1}{C'})^2}{16\lambda^2(1+|2p-1|)^4}\frac{1}{\alpha}N
\end{aligned}
$$

with $\gamma < 1$ and close to 1 (as ε becomes small). The right hand side of this inequality becomes negative when δ and α become small appropriately. To check whether this can be done in agreement with (12) we insert

$$
\alpha = \sqrt{\delta(1-\delta)}\frac{1}{\sqrt{3}\lambda(1+|2p-1|)^2}(1-(2p-1)^2+\frac{1}{C'})(1-|2p-1|)
$$

into the right hand side of (13) and obtain

$$
\left(-\frac{\sqrt{3}\gamma(1-(2p-1)^2+\frac{1}{C'}}{16\lambda(1+|2p-1|)^2}(1-|2p-1|)\sqrt{\delta(1-\delta)}-\delta\log\delta-(1-\delta)\log(1-\delta)\right)N. \qquad (14)
$$

As it is quickly checked that for each positive constant C there is an interval $[0,r]$ (depending on C, of course) such that

$$
C\sqrt{\delta(1-\delta)} \geq -\delta\log\delta-(1-\delta)\log(1-\delta)
$$

for all $\delta \in [0,r]$, the above exponent becomes negative if we choose δ small enough and e.g. α as the right hand side of (12). This completes the proof of the theorem. □

Let us finally comment a little on the result of Theorem 5. Observe that the bound on the moment generating function in (14) as well as the bound on α in (13) depends on p mainly via the factor $(1-|2p-1|)$ (the other terms containing p are bounded from above and away from 0) which converges to zero for p close to one or close to zero and therefore can only deteriorate the bounds for α (allowing smaller $\alpha's$ only) for large correlations. Due to the many estimates in the proof of Theorem 5 this is, of course, in no way a proof that the storage capacity decreases with an increasing correlation (only our bounds do), but it might either indicate that the Hopfield model has problems to store patterns with large correlations or it just shows that our estimates get worse for large p (which is probably true). However, as already mentioned after Theorem 4, a decrease of storage capacity (when the correlation increases) would not be totally unexpected due to the way the Hopfield model is assumed to work. On the other hand from the point of view of information theory, sequence of correlated data contains less information than an independent sequence (e.g. in the extreme case that all patterns agree it suffices to know the first patterns to reconstruct them all). Hence one could expect a reasonable neural network to be able to store more correlated patterns than uncorrelated ones. Indeed, as shown in [Lö96a], provided we know the p of our Markov chain and therefore the

covariance of the patterns in advance (note that we do not impose to know the empirical correlations), there exists a variant of the Hopfield model that can store a larger number of correlated data than the number of independent patterns one can store in the standard Hopfield model provided the first notion of storage capacity is used. With the second notion of storage capacity a bound of αN with α not depending on p is obtained.

4 The Storage Capacity of the Hopfield Model with Biased Patterns

Finally we will briefly report on some recent results on the storage capacity for the Hopfield model with biased patterns obtained in [Lö96b]. More precisely we will assume that the patterns are i.i.d. as in Section 2 but have a uniform bias, i.e.

$$P(\xi_i^\nu = 1) = p \quad \text{and} \quad P(\xi_i^\nu = -1) = 1 - p. \tag{15}$$

As already pointed out several times in the physical literature (see e.g [HK91]) the standard Hopfield model as introduced above cannot store any increasing amount of such patterns, simply because the local field associated with the Hopfield Hamiltonian h_i^μ at site i and for a pattern μ

$$h_i^\mu := \xi_i^\mu + \sum_{\substack{\nu \neq \mu \\ j \neq i}} \xi_j^\mu \xi_i^\nu \xi_j^\nu$$

quickly gets dominated by the bias from the second term for $M \to \infty$. To overcome this difficulty we center the patterns in the Hamiltonian, i.e. we consider synaptic efficacies of the form

$$J_{ij} = \sum_{\nu=1}^{M(N)} \overline{\xi_i^\nu}\, \overline{\xi_j^\nu},$$

where
The $\overline{\xi_i^\nu}$ are the centered patterns ξ_i^ν, i.e.

$$\overline{\xi_i^\nu} = \xi_i^\nu - (2p - 1).$$

This leads to the Hamiltonian of the biased Hopfield model

$$\overline{H_N(\sigma)} = -\frac{1}{2N} \sum_{i,j=1}^{N} \sigma_i \sigma_j J_{ij} = -\frac{1}{2N} \sum_{i,j=1}^{N} \sum_{\nu=1}^{M(N)} \sigma_i \sigma_j \overline{\xi_i^\nu}\, \overline{\xi_j^\nu}. \tag{16}$$

For this variant of the Hopfield model we have the following results

Theorem 6 *Assume the random patterns ξ^ν fulfill (15) and $M(N) = \frac{N}{\gamma \log N}$. Then for the Hopfield model (16) the following assertions hold true:*

1. *If $\gamma > \frac{3}{8p^2(1-p)^2}$*

$$P(\liminf_{N \to \infty} (\cap_{\nu=1}^{M(N)} T\xi^\nu = \xi^\nu)) = 1$$

i.e. the patterns are almost surely stable.

2. *If* $\gamma > \frac{1}{4p^2(1-p)^2}$

$$P((\cap_{\nu=1}^{M(N)} T\xi^\nu = \xi^\nu)) = 1 - R_N$$

with $\lim_{N\to\infty} R_N = 0$.

3. *If* $\gamma > \frac{1}{8p^2(1-p)2}$ *for every fixed* $\nu = 1, \ldots, M$

$$P(T\xi^\nu = \xi^\nu) = 1 - R_N$$

with $\lim_{N\to\infty} R_N = 0$.

Here, of course, T *is the gradient dynamics defined as in Section 1 for the Hamiltonian (16).*

Note that the estimates of the above Theorems for $p = \frac{1}{2}$ (the unbiased case) agree with the results in the standard Hopfield model. I may of course be true that the estimates can be improved in some respects. Note however, that our bound on the storage capacity of the Hopfield model with biased patterns is (similar to the case of correlated patterns) a decreasing function in the bias of the patterns.

We now give a result on the storage capacity of the Hopfield model with biased patterns provided that Newman's concept of storage is used. It turns out that a bias does not destroy the storage abilities of the Hopfield model and that it can store "extensively many" patterns (i.e. M(N) grows like αN), although the critical α decreases to zero when the bias get s large.

Theorem 7 *Suppose that the random patterns fulfill (15). There exists an* $\alpha_c > 0$ *(depending on p) such that if* $M(N) \leq \alpha_c N$, *then there are* $\varepsilon > 0$ *and* $0 < \delta < 1/2$ *such that for the standard Hopfield model (16)*

$$P\left(\liminf_{N\to\infty}(\cap_{\nu=1}^{M(N)} \cap_{\sigma\in S_\delta(\xi^\nu)} (\overline{H_N(\sigma)} \geq \overline{H_N(\xi^\nu)} + \varepsilon N))\right) = 1$$

where $S_\delta(\xi^\nu)$ *is the Hamming sphere of radius* δN *centered in* ξ^ν.

Note that these results resemble the results of the Hopfield model with correlated patterns obtained in [Lö96a].

A proof of the above theorems can be carried out along the ideas introduced in the proofs of Theorems 4 and 5 (and uses nearly the same inequalities). The interested reader may consult [Lö96b] for details.

References

[AGS87] D.J. Amit, G. Gutfreund, H. Sompolinsky; Statistical mechanics of neural networks near saturation; Ann. Phys. 173, 30-67 (1987)

[BG92] A. Bovier, V. Gayrard; Rigorous bounds on the storage capacity of the dilute Hopfield model; J. Stat. Phys. 69, 597-627 (1992)

[Bu94] D. Burshtein; Nondirect convergence radius and number of iterations of the Hopfield associative memory; IEEE Trans. Inf. Th. 40, 838-847, (1994)

[HK91] L. van Hemmen, R. Kühn, Collective phenomena in neural networks; in Models of neural networks, E. Domany, L. v. Hemmen, R. Schulte (eds.), Springer, Berlin (1991)

[Ho82] J.J. Hopfield; Neural networks and physical systems with emergent collective computational abilities; Proc. nat. Acad. Sci. USA 79, 2554-2558 (1982)

[KP88] J.Komlos, R. Paturi; Convergence results in an associative memory model; Neural Networks 1, 239-250 (1988)

[Lö96a] M. Löwe; On the storage capacity of Hopfield models with weakly correlated patterns, Preprint, Universität Bielefeld, submitted (1996)

[Lö96b] M. Löwe; On the storage capacity of the Hopfield model with biased patterns, Preprint, Universität Bielefeld, submitted (1996)

[L94] D. Loukianova; Capacité de mémoire dans le modèle de Hopfield; C.R. Acad. Sci. Paris 318, 157-160 (1994)

[MPRV87] R. McEliece, E. Posner, E. Rodemich. S. Venkatesh; The capacity of the Hopfield associative memory; IEEE Trans. Inf. Th. 33, 461-482 (1987)

[N88] C. Newman; Memory capacity in neural networks; Neural Networks 1, 223-238 (1988)

[P95] D. Petritis; Thermodynamic formalism of neural computing; Preprint, Université de Rennes I (1995)

[T96] M. Talagrand; Rigorous Results of the Hopfield Model with Many Patterns; Preprint, Paris (1996)